Proceedings of the 2nd International Conference on Biological Engineering and Medical Science (ICBioMed 2022), Part I
DOI: 10.54254/2753-8818/3/20220351

Heart Disease Diagnosis Associated with Potential Causative Factors Based on the 2020 Data

Ruiyu Wang

University of Toronto, Toronto, ON, Canada

Jasminery.wang@mail.utoronto.ca

Abstract. Heart disease is one of the leading causes of death all over the world. In order to determine the relationships between heart disease and some potential causing factors, this paper conducted the data analysis based on the survey results from the Centres for Disease Control and Prevention (CDC). By analyzing the correlation index and p-index, the author defined the general health condition, age, and background disease as the potential causative causing factors. At the same time, a logic regression model is applied to explore the relationship between them. It could be found as a result that with age increasing, people will face a higher risk of suffering from heart disease. Also, unhealthy daily life habits (i.e., smoking) would also increase the risk of heart disease. Furthermore, background diseases such as diabetics could also be a potential causative factor.

Keyword: heart disease, association, centres for disease control and prevention, logistic regression, correlation index, P-index, diabetes, kidney disease

1. Introduction

During the past decade, ten causes have resulted in more than 70% of all deaths in America. Heart disease has been claimed as the first leading cause of death, responsible for approximately 23.5% of total deaths [1][2]. Heart disease can be classified into four types based on its underlying cause: coronary artery and vascular disease, heart rhythm disorders, structural heart disease, and heart failure [3]. Usually, an electrocardiogram(EKG) will be used as the first test, by recording electrical activity within the heart. Besides, other tests (i.e. cardiac catheterization, X-rays) would also be applied to heart disease detection [4][5].

In this paper, we will analyze the dataset conducted by the Centres for Disease Control and Prevention (CDC) in 2020 by annual phone surveys from more than 400,000 adults related to their health status. The interview questions include gender, age, race, living habits (i.e. smoking, alcohol drinking) and physical health (i.e. general physical health and other diseases) [6].

In the data analysis of this paper, the author used some basic model tests, such as logistic analysis, etc., by analyzing the correlation index and p-index, to explore the relativeness between the relationship between heart disease and different factors, including the basic information(age, gender, and race), daily life habits, and other health conditions, to identify and discuss some common heart disease triggers.

By processing and analyzing this data set, we will derive the potential impact of different factors on heart disease and explore the possible consequences of their combinations. This result is expected to be used for the prevention and prediction of heart disease to avoid the possible risk of heart disease.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/3/20220351

2. Methodology

2.1. Study design and participants

As mentioned above, the Centres for Disease Control and Prevention (CDC) conducts the annual phone survey and is a major part of the Behavioural Risk Factor Surveillance System (BRFSS). It is described on the official website as "Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world" [7]. The dataset used in the memo is the most recent data from February 2022, including the data from 2020.

2.2. Questionnaire

The questionnaire includes the status of heart disease, BMI value, gender, age, race, sleeping time, the status of general health, the daily-life habits regarding smoking, alcohol drinking, and physical activity, and whether the interviewer has stroke, difficulty walking, diabetic, asthma, kidney disease, or skin cancer.

In terms of the age category, except for the group of 18-24 and 80 or older, the in-between ages is divided into five-year-old intervals. Furthermore, the general health category was divided into five groups, which were excellent, very good, good, fair, and poor, respectively.

3. Statistical analysis

3.1. Data selection and exclusion

The statistical analyses were conducted by Python (version 3.10) and SPSS.

In order to ensure the accuracy and precision of the sample and to exclude unnecessary interference, the group of people who are aged 80 and older has been excluded. After the exclusion, a total number of 295,642 samples were included and analyzed in the memo. Moreover, when exploring the logistic regression relationship between the potential causative factors and the risk of heart disease, since the total number of samples was too large, we decided to randomly select 30,000 individuals for the analysis. To ensure the rigor of the data, we used the random function in Microsoft Excel. Figure 1 below shows the process of the data inclusion algorithm.

DOI: 10.54254/2753-8818/3/20220351

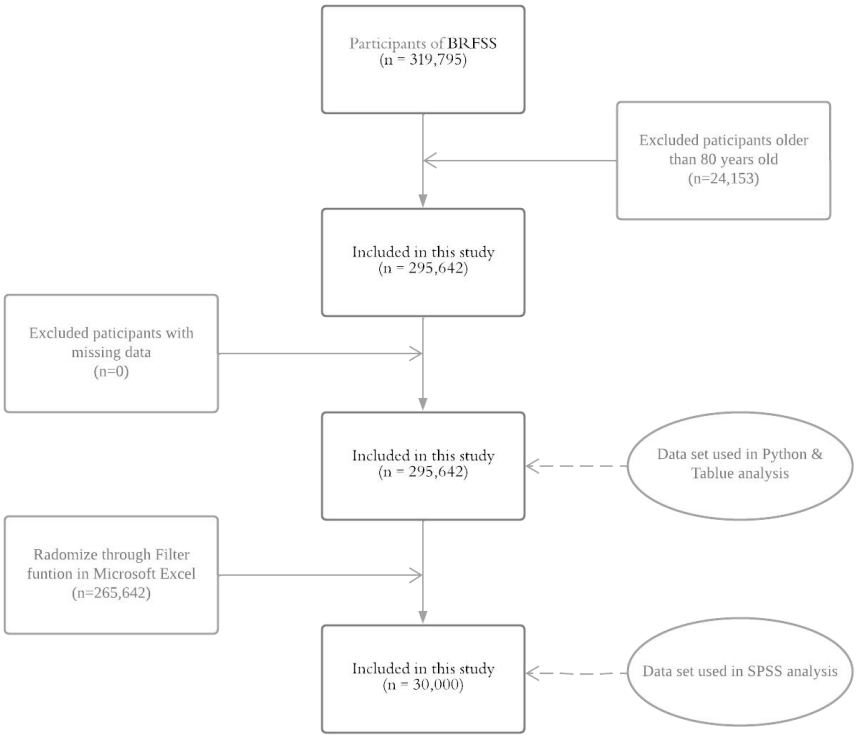


Figure 1. Inclusion Algorithm.

3.2. The basic association between each potential reason & heart disease

First of all, we collated the associations between potential factors and whether respondents suffered from heart disease. It could be seen clearly that among the heart disease patients, around 58.6% of them have the habit of smoking (see Figure 2). Meanwhile, we found that among respondents with heart disease, the gender ratio was not approximately 1:1 but reached almost 3:2 (the ratio of male heat disease patients is 59%). It indicates a higher potential risk of heart disease for men than women (see Figure 3). Moreover, a significant increasing trend in the risk of heart disease would be observed with the change of age (see Figure 4).

DOI: 10.54254/2753-8818/3/20220351

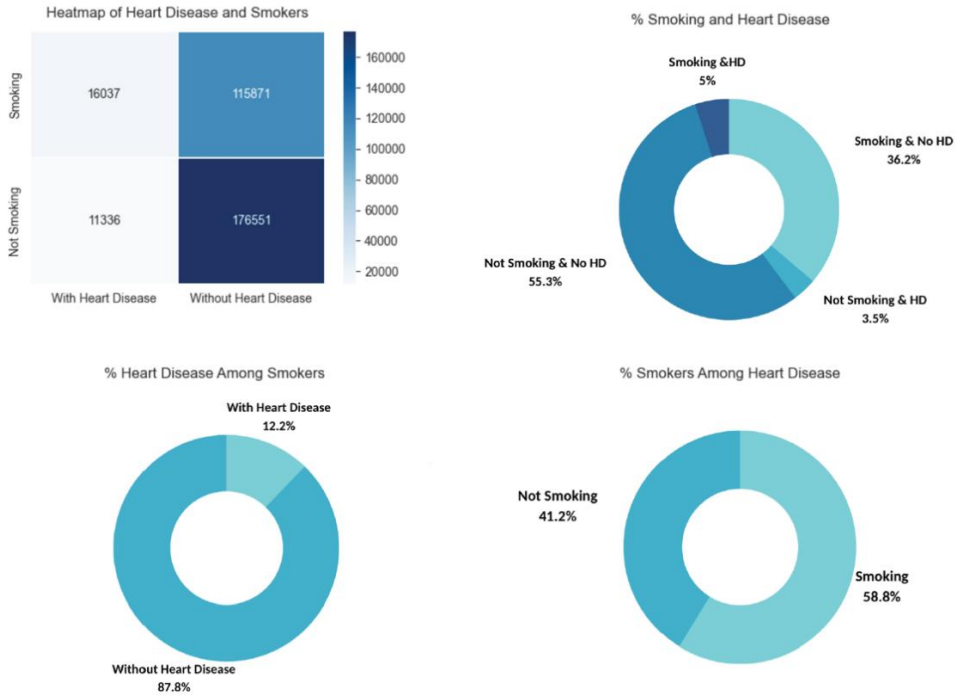


Figure 2. Relationship between smoking habit and heart disease.

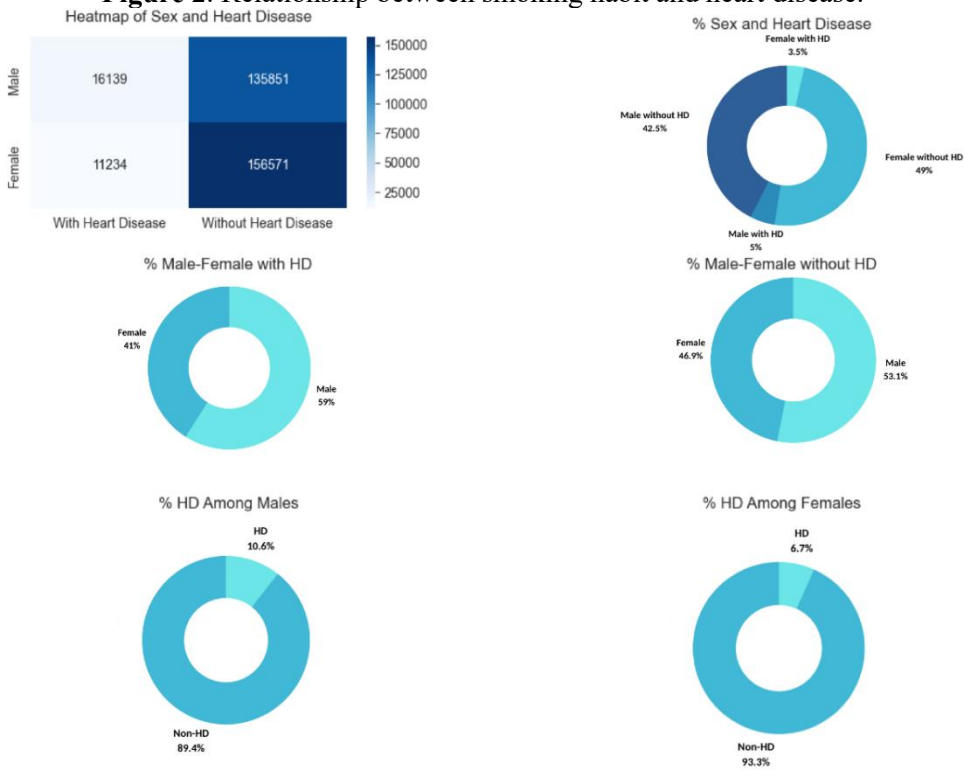


Figure 3. Relationship between gender and heart disease.

DOI: 10.54254/2753-8818/3/20220351

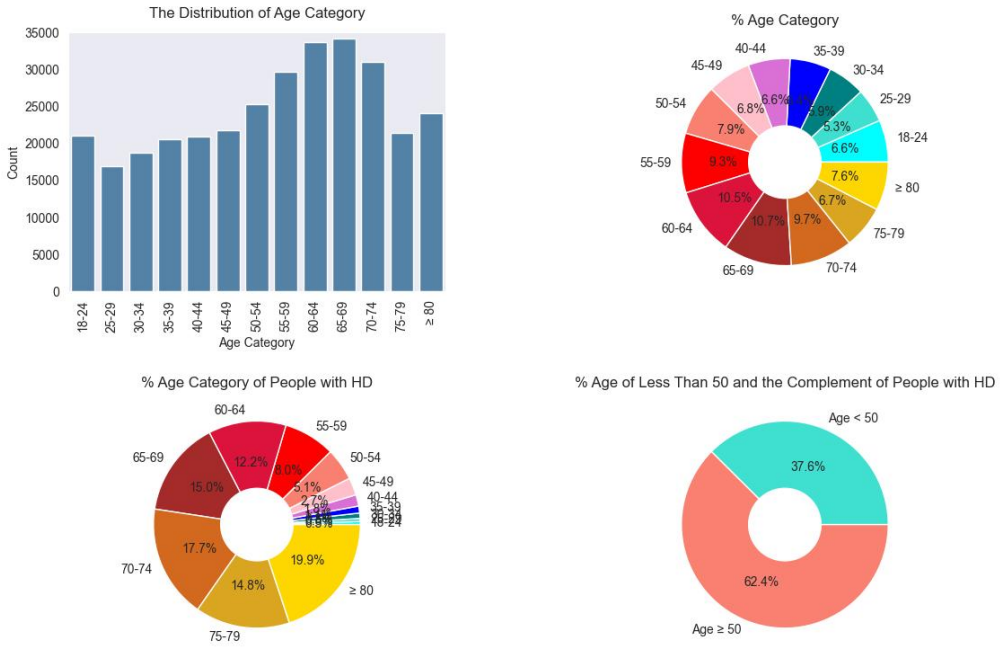


Figure 4. Relationship between age and heart disease.

3.3. Correlation coefficient and p-value

A correlation coefficient value and a significant P value are the most common ways to establish the relationship between variables. The correlation relationship could test the existence of the relationship between two variables, while the p-value is able to show the level of statistically significant [8]. Usually, a p-value less than 0.05 can be valid proof of a correlation.

The table below shows the model test results between heart disease and some potential factors. The p-values of BMI value (p-value = 0.00), and sleep time (p-value = 0.043) with heart disease reveal the statistical significance. The result of the data analysis is the same as the trend predicted from the charts mentioned before (see Figure 5), which proves the correctness of the previous speculation. Meanwhile, the p-value between general health status and heart disease is less than 0.05 and demonstrates statistical significance.

DOI: 10.54254/2753-8818/3/20220351

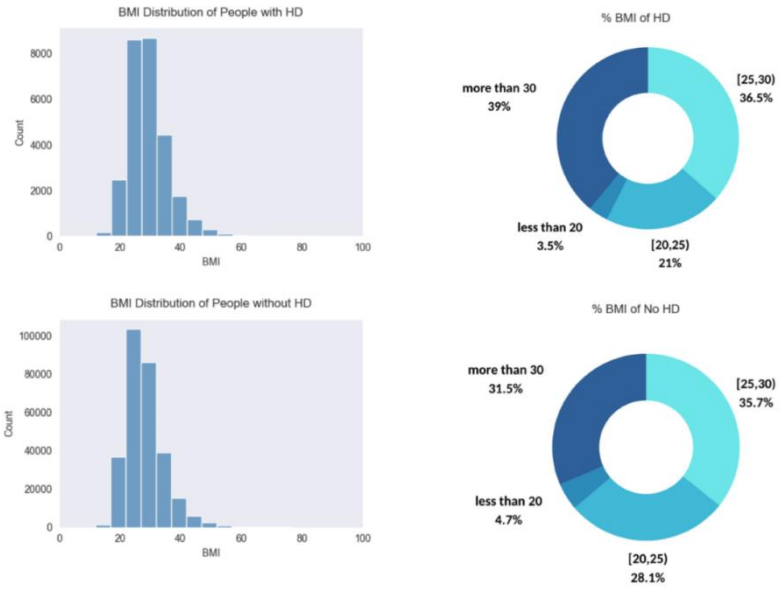


Figure 5. Relationship between BMI value and heart disease.

Table 1. Correlation coefficient table (heart disease with BMI, sleep time, general health and gender).

	Heart Disease	BMI	Sleep Time	General Health	Gender
Heart Disease	1.000(0.000***)	0.047(0.000***)	0.012(0.043**)	0.021(0.000***)	-0.067(0.000***)
BMI	0.047(0.000***)	1.000(0.000***)	-0.047(0.000***)	-0.115(0.000***)	-0.033(0.000***)
Sleep Time	0.012(0.043**)	-0.047(0.000***)	1.000(0.000***)	-0.004(0.460)	0.006(0.276)
General Health	0.021(0.000***)	-0.115(0.000***)	-0.004(0.460)	1.000(0.000***)	-0.002(0.771)
Gender	-0.067(0.000***)	-0.033(0.000***)	0.006(0.276)	-0.002(0.771)	1.000(0.000***)

Note: ***, **, * represent the significance levels of 1%, 5%, and 10%, respectively.

When it comes to the relationship between heart disease and background disease, we can find that stroke and diabetes would not be defined as potential factors, with correlation coefficients of 0.196 and 0.123 and p-values of 0.00 and 0.00, respectively. From the model test, there is a particular relationship between whether there is a heart disease and the interviewer's previous disease history. This may be because the underlying disease can lead to hormonal changes and a decline in physical health, leading to a greater susceptibility to the underlying disease and vice versa.

In addition, the risk of heart disease increases with age. At the same time, it can also be seen from the table that increasing age will also lead to an increased risk of other diseases(see the underlined text).

Table 2. Correlation coefficient table (heart disease with stroke, diabetic and age).

	Heart Disease	Stroke	Diabetic	Age Category
Heart Disease	1.000(0.000***)	0.196(0.000***)	0.123(0.000***)	-0.004(0.451)
Stroke	0.196(0.000***)	1.000(0.000***)	0.075(0.000***)	0.014(0.015**)
Diabetic	0.123(0.000***)	0.075(0.000***)	1.000(0.000***)	0.017(0.003***)

DOI: 10.54254/2753-8818/3/20220351

	Heart Disease	Stroke	Diabetic	Age Category
Age Category	-0.004(0.451)	0.014(0.015**)	0.017(0.003***)	1.000(0.000***)

Note: ***, **, * represent the significance levels of 1%, 5%, and 10%, respectively.

3.4. Logistic regression

A logistic regression is usually used to predict a categorical variable. In this demo, since there are many factors that may cause heart disease, and there may be multiple factors acting at the same time, we will use logistic regression to judge and predict the impact of multiple factors on heart disease.

Table 3. Binary logistic regression for heart disease and some factors (continue).

item	Regression coefficients	standard error	Wald	p-value	OR value	95% confidence interval for the OR value	
						upper limit	lower limit
constant	-3.309	0.044	5675.635	0.000***	0.037	0.034	0.04
Mental Health	0.007	0.003	7.612	0.006***	1.007	1.002	1.012
Smoking_Yes	0.714	0.043	272.337	0.000***	2.043	1.877	2.224
Sex_Male	0.462	0.044	112.889	0.000***	1.588	1.458	1.729
Diabetic2_Yes	1.106	0.048	538.889	0.000***	3.022	2.753	3.318
Kidney Disease_Yes	1.316	0.072	331.797	0.000***	3.729	3.237	4.297
Dependent variable: Heart Disease							

Note: ***, **, * represent the significance levels of 1%, 5%, and 10%, respectively.

Based on current research results (discussed in later sections), the leading causes of diabetes are smoking, high blood lipids, and high blood pressure[9]. In connection with the conclusions obtained from the previous observation of the chart, we chose to use several factors such as mental health (which may lead to changes in hormone levels), smoking habits, gender, and underlying diseases of diabetes and kidney disease as independent variables to explore the relationship between these factors and heart disease risk.

From the table above, all these factors are positively related to the risk of heart disease. Among them, the correlation indices of mental health status and gender for heart disease risk were 0.007 and 0.462, respectively. The correlation index for smoking habits and heart disease was 0.714, while the correlation indices for diabetes and kidney disease were 1.106 and 1.316, respectively. It is worth mentioning that diabetes can cause blood sugar levels to rise, and diseased kidneys are less able to help regulate blood pressure, which can lead to increased blood pressure. Therefore, although diabetes and kidney disease are not directly linked to heart disease, the impact of such underlying diseases on the body can lead to an increased risk of heart disease.

Table 4. Classification evaluation index.

Accuracy	recall	precision	F1	AUC
0.914	0.914	0.88	0.877	0.697

The classification evaluation indicators are shown in the above table, and quantitative indicators further measure the classification effect of logistic regression. It can be seen that the accuracy and accuracy of this model are high, and the proportion of correct predicted samples in the total samples is large. At the same time, the AUC value of this model is close to 1, indicating that this model has a relatively good classification correlation.

DOI: 10.54254/2753-8818/3/20220351

4. Discussion

4.1. Literature review

In order to confirm the reliability of the analysis results, we compared some recent scientific literature with the existing data analysis results. High levels of blood sugar and blood pressure have been identified as significant causes of heart disease [10]. Also, smoking has been shown to increase the risk of heart disease. However, no more authoritative data proves that men are at greater risk of heart disease than women. This may be because, in the current society, more men have bad habits such as smoking, which leads to a higher proportion of men with heart disease among the respondents.

4.2. Strength & limitations

This study has distinct advantages and disadvantages. First of all, the data source is very authoritative (CDC), and the content of the questionnaire is relatively detailed. Therefore, our analysis can analyze as many potential links between various factors and heart disease risk as possible. At the same time, the sample size is considerable (more than 300,000 respondents). Therefore, even if some interference items are excluded, the remaining analyzable sample size is still considerable, avoiding the problem of unrepresentative conclusions due to the small sample size.

However, it must be acknowledged that this analysis's limitations are also undeniable. First of all, due to the limitation of professional knowledge, this analysis is not too in-depth but only stays at the stage of basic model testing. Also, since most of the data source's independent variables are categorical, the statistical models that can be tested are relatively limited.

5. Conclusion

To sum it up, there are a number of factors that may increase the risk of heart disease. Among them, in addition to increasing age, unhealthy lifestyle habits (smoking) can also lead to an increased risk of disease. Also, some diseases that cause changes in hormone levels in the body, or other health indicators, may also be the main cause of heart disease. However, it has to be acknowledged that the data analysis has obvious limitations. Only simple model test is used in the analysis, and the combined impact of the potential causative factors has not been discussed. In order to better improve the data analysis model and make up for the current deficiencies, the author hopes to take the following methods in the future. More advanced and accurate models are expected to be applied to the analysis to obtain more comprehensive analysis and prediction results. At the same time, more factors and combinations of factors will also be discussed. The author looks forward to learning more in-depth data analysis knowledge, disease and epidemic prevention and control knowledge, and applying them in the future.

References

[1] MediLexicon International. (n.d.). The top 10 leading causes of death in the United States. Medical News Today. Retrieved August 1, 2022, from https://www.medicalnewstoday.com/articles/282929#heart-disease

[2] World Health Organization. (n.d.). Cardiovascular diseases (cvds). World Health Organization. Retrieved August 30, 2022, from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[3] MediLexicon International. (n.d.). Cardiovascular disease: Types, symptoms, prevention, and causes. Medical News Today. Retrieved August 30, 2022, from https://www.medicalnewstoday.com/articles/257484

[4] WebMD. (n.d.). Understanding heart disease. WebMD. Retrieved August 1, 2022, from https://www.webmd.com/heart-disease/understanding-heart-disease-symptoms

[5] MediLexicon International. (n.d.). Aspirin: Health benefits, uses, risks, and side effects. Medical News Today. Retrieved August 30, 2022, from https://www.medicalnewstoday.com/articles/161255

DOI: 10.54254/2753-8818/3/20220351

[6] Pytlak, K. (2022, February 16). Personal key indicators of heart disease. Kaggle. Retrieved August 1, 2022, from https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

[7] Centers for Disease Control and Prevention. (2019, December 9). Know your risk for heart disease. Centers for Disease Control and Prevention. Retrieved August 1, 2022, from https://www.cdc.gov/heartdisease/risk_factors.htm

[8] Correlation and P value. The Data School. (n.d.). Retrieved August 1, 2022, from https://dataschool.com/fundamentals-of-analysis/correlation-and-p-value/#:~:text=The%20two%20most%20commonly%20used,an%20experiment%20is%20statistically%20significant.

[9] Heart disease and stroke. Heart Disease and Stroke | Healthy People 2020. (n.d.). Retrieved August 1, 2022, from https://www.healthypeople.gov/2020/topics-objectives/topic/heart-disease-and-stroke#:~:text=High%20blood%20pressure%2C%20cigarette%20smoking,national%20epidemic%20of%20cardiovascular%20disease

[10] Goldstein, S. A., D'Ottavio, A., Spears, T., Chiswell, K., Hartman, R. J., Krasuski, R. A., Kemper, A. R., Meyer, R. E., Hoffman, T. M., Walsh, M. J., Sang, C. J., Paolillo, J., & Li, J. S. (2020). Causes of death and cardiovascular comorbidities in adults with congenital heart disease. *Journal of the American Heart Association*, **9(14)**. https://doi.org/10.1161/jaha.119.016400